Assignment-2

Submitted by: PEDDAPURAM SHIVA KUMAR – 811235874

submitted to: - Jun Li

subject: - Applied Statistics (MATH 50015-001)

**2.2 (Data file: UBSprices)**

**The international bank UBS regularly produces a report (UBS, 2009) on prices and earnings in major cities throughout the world. Three of the measures they include are prices of basic commodities, namely 1kg of rice, a 1kg loaf of bread, and the price of a Big Mac hamburger at McDonalds. An interesting feature of the prices they report is that prices are measured in the minutes of labour required for a “typical” worker in that location to earn enough money to purchase the commodity. Using minutes of labour corrects at least in part for currency fluctuations, prevailing wage rates, and local prices. The data file includes measurements for rice, bread, and Big Mac prices from the 2003 and the 2009 reports. The year 2003 was before the major recession hit much of the world around 2006, and the year 2009 may reflect changes in prices due to the recession. The figure below is the plot of y = rice2009 versus x = rice2003, the price of rice in 2009 and 2003, respectively, with the cities corresponding to a few of the points mark**

Chart, scatter chart

Description automatically generated

**2.2.1 The line with equation y = x is shown on this plot as the solid line. What is the key difference between points above this line and points below the line?**

**Explanation: -**

From the above plot we can say that line Y=X determines the data points that are on the line.

It also concludes that the rice prices over different years are the same. We can also conclude that the data in the lower limits has less expected errors. Most of the data in the lower that is Seoul is far away from the ols and y=x line.

Another key observation is Nairobi point is meeting as per our expectation that is on ols.

Mumbai is away from the both the lines i.e., from y=x and ols line which states that the dataset for the Mumbai city is not expected to be correct as per the graph it is throwing lot of error.

Finally, we can conclude that points that are above the line I.e., x=y, indicate that the price of rice in 2009 is higher when compared to the price in 2003.

The points that are below the line say that the price of rice is less in 2009 when compared to the year 2003.

**2.2.2 Which city had the largest increase in rice price? Which had the largest decrease in rice price?**

**Explanation:**

from the above graph we can see that Vilnius is the city that has the largest increase in rice prices and Budapest is in second lead... and coming to the point of decrease in the price Mumbai Seoul ranks better

**2.2.3** **The ols line ˆ ˆ ˆ y x = + β β 0 1 is shown on the figure as a dashed line, and** **evidently ˆ β1 < 1. Does this suggest that prices are lower in 2009 than in 2003? Explain your answer.**

**Answer: -**

From the plot, we can say that prices are not less in 2009 because most of the points lie above the line where we can conclude that prices have been higher in 2009 when compared to 2003.

**2.2.4: - Give two reasons why fitting simple linear regression to the figure in this problem is not likely to be appropriate.**

**Explanation:**

1.) The main reason why simple linear regression is not suitable for the above graph is basically in simple linear regression we use only one factor/variable that can be predicted but here we have different cities which have different parameters. So, when we take city like Mumbai Nairobi Vilnius which are getting impacted, so these terms are ultimately becoming errors terms as they are not on the OLS estimator.

2.) But these cities also have some data which can support/ follows ols estimator if we use any other plot model.

3.) the major issue is with the predictor in the above graph. Because we can clearly see that there are lot of errors when compared to the actual’s values and for the cities like Mumbai they are not even well expressed.

So, considering all the above points we can conclude that simple linear regression model is a problem in the above figure as it is not obeying many points and it is not making a best part in graphing.

**2.3) (Data file: UBSprices) This is a continuation of Problem 2.2. An alternative representation of the data used in the last problem is to use log scales, as in the following figure:**

Chart, scatter chart

Description automatically generated

**2.3.1 Explain why this graph and the graph in Problem 2.2 suggests that using log-scale is preferable if fitting simple linear regression is desired.**

**Answer: -**

In General, by using logarithms functions we reduce the errors.

So, in simple linear regression we mainly focus on logarithms function so that we can get reduced with errors and can directly focus on the proper plotting. The distribution of the points on the axes is no longer skewed and variability appears constant.

When we see the graph y=x i.e., 2.2 we see that there are lot of errors (the points which are away from the ols parameter are considered as errors) and the data points are not in a properly visible.

Now, in the figure above when we used logarithm function, we can clearly see almost all the points at the extreme right are no longer separated from the other points. Vilnius and Budapest still appear to be outliers and coming to the points of errors values we can see the error values are reduced when compared with 2.2.

We can also observe that in the log graphs we can clearly see lot of data points that are spread around the ols parameter

**2.3.2) Suppose we start with a proposed model “E | ( ) y x = γ xβ”**  
 **This is a common model in many areas of study. Examples include allometry (Gould, 1966), where x could represent the size of one body characteristic such as total weight and y represents some other body characteristic, such as brain weight, psychophysics (Stevens, 1966), in which x is a physical stimulus and y is a psychological response to it, or in economics, where x could represent inputs and y outputs, where this relationship is often called a Cobb– Douglas production function (Greene, 2003). If we take the logs of both sides of the last equation, we get log(E | ( ) y x ) l = + og( ) γ β 0 1 log(x) If we approximate log(E(y|x)) ≈ E(log(y)|x), and write β0 = log(γ), to the extent that the logarithm of the expectation equals the expectation of the logarithm, we have E | (log( )y x) l = + β β 0 1 og( ) x Give an interpretation of β0 and β1 in this setting, assuming β1 > 0**.

**Answer: -**

For the above-mentioned proposed model of beta1 the interception is going to growth in exponentially if beta1 value is greater than 1 i.e., (b1>1)

And if the beta1 value is equal to 1 then we see a linear growth.

If beta1 is less than 1 decreases.

If gamma0 > 1, beta0 > 0 the curve will shift up

if beta0<0 then the curve is shifted down.

**2.4 (Data file: UBSprices) This problem continues with the data file UBSprices described in Problem 2 .2**

**2.4.1 Draw the plot of y = bigmac2009 versus x = bigmac2003, the price of a Big Mac hamburger in 2009 and 2003. On this plot draw (1) the ols fitted line; (2) the line y = x. Identify the most unusual cases and describe why they are unusual.**

Code: - Text

Description automatically generated

Normal plotting without any lines

Chart, scatter chart

Description automatically generated

Final plot with ols Fitted line, line with y=x.

Chart, scatter chart

Description automatically generated

The above figure shows the plot between the prices of bigmac2009 versus the bigmac2003.

Blue line represents the OLS Fitted line.

Green line indicates the line y=x.

Also, as we observe from the above plot most of the data points lie closer to the fitted line and one point lies far away from the other data points which we can say is an unusual case.

**2.4.2 Give two reasons why fitting simple linear regression to the figure in this problem is not likely to be appropriate.**

**Answer: -**

first reason is that there is not so much correlation between these variables.

data variability is less.

**2.4.3) Plot log(bigmac2009) versus log(bigmac2003) and explain why this graph is more sensibly summarized with a linear regression.**

code

A picture containing text

Description automatically generated

Chart, scatter chart

Description automatically generated

**Explanation: -**

The above plot shows the logarithmic prices of bigmac2009 versus bigmac2003. There are two lines mentioned in the graph

1. The black line indicates the fitted line of log(bigmac2009) vs log(bigmac2003),
2. the red line indicates the line y=x (which we have used in 2.2,2.3 problems).

From the above graph we can see that the data points are properly fitted when we have used logarithmic function. Moreover, it is sensibly summarized with a proper linear regression, we can conclude that the errors value got reduced when compared with other graphs. So, we can finally conclude that by using logarithmic transformations/function we get the best output, or our model gets efficiently fit.